7 Appendix

This supplementary material includes the results of ablation study on 3 important components of the proposed methodology and also mentions about the code which can be used to reproduce the results as reported in this paper.

8 Ablation Study (continued...)

8.1 Effect of the Attention-ACM module

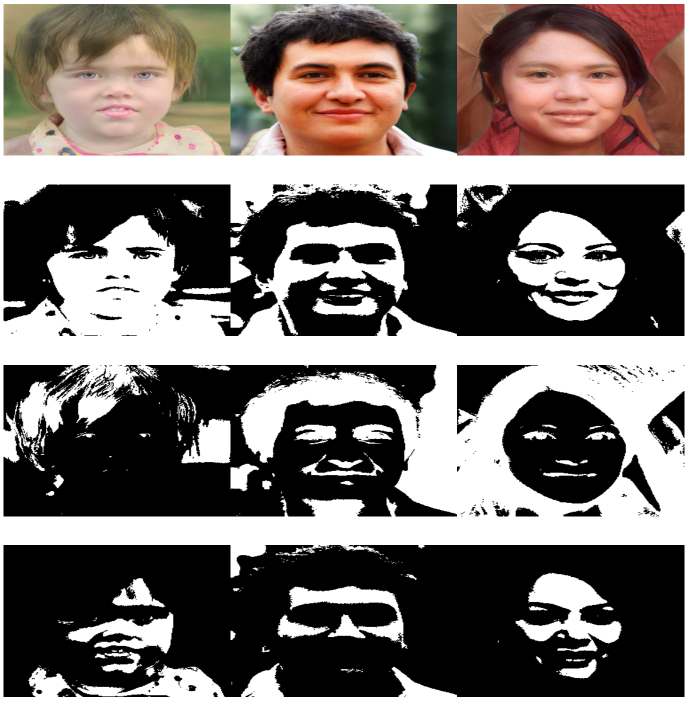


Fig. 4. The above images are a few examples generated corresponding different descriptions. Each row represents a feature map taken from the attention affine combination module. The first row represents the generated images, following rows show the feature maps of $W(v)$, $b(v)$ and $h \odot W(v)$, respectively. The attention and Hadamard product enables text representations h to re-weight image feature maps, which serves as a regional selection purpose to help the model precisely identify desired attributes matching the given text, and in the meantime the correlation between attributes and semantic words is built for effective feature generation and manipulation.

8.2 Effect of the Trained Latent Space

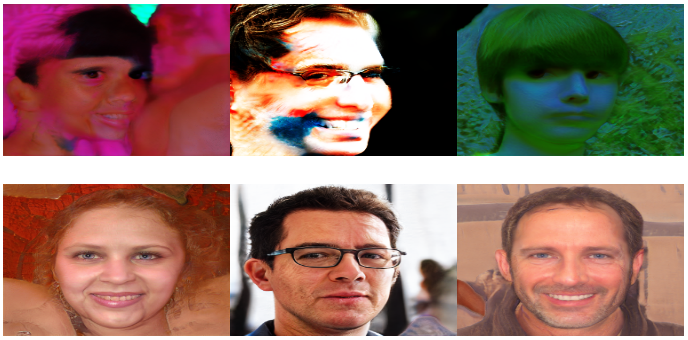
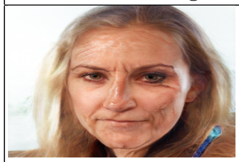
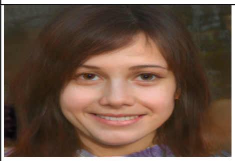
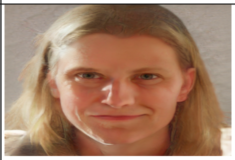
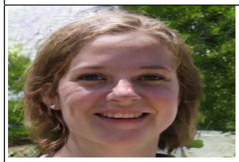
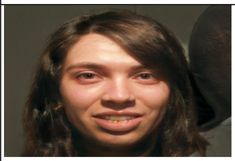
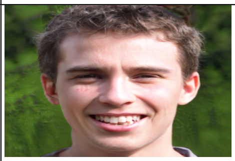


Fig. 5. The above images are a few examples generated without and with our trained latent space, respectively. In case of the results generated without the latent space, the noise vector generated from the cyclic network is directly fed into the generator without an forward pass along the encoder-decoder module.

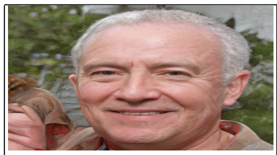
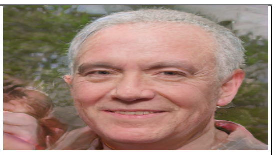
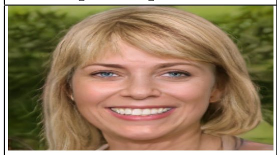
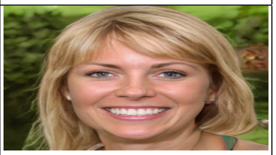
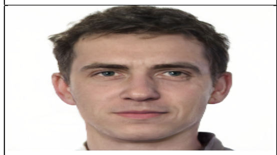
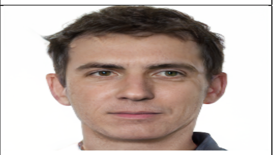
8.3 Effect of the Cyclic Network

Table 6. A table showing some text inputs and their corresponding image results. The first row consists of the images that are generated by feeding the generator with text and a randomly sampled noise vector. The second row of images consists of the images generated by replacing the randomized vectors with the vectors generated by the cyclic network. As the images suggest, cyclic network significantly improves the accuracy of features, which is due to the ability of the network to learn a latent representation of an image, solely based on the input description.

This is a woman. She has medium length blonde hair. She is smiling and is a little old. She has a big nose.	She is a young woman with an oval face. Heir hair is long and brown. She has a broad forehead.	This is a young man with brown hair. He is smiling and has narrow eyes. The man has bushy eyebrows.
		
		

8.4 Significance of BERT

Table 7. Given the same input noise vector, our model generates similar images from descriptive inputs and mentions the attributes present. This proves that despite training the model on the CelebA [20] dataset using comma-separated attributes as labels, with the help of the BERT [2] network, our model is semantically able to understand the input sentences and generate the corresponding attributes.

 This man is smiling. He is old and has short white hair. He has a receding hairline. His nose is sharp and pointed.	 Old, male, short white hair, pointed nose, smiling, receding hairline
 This woman is young and smiling. Her hair is blond. She has high cheekbones. The woman has a broad forehead, and her eyebrows are arched.	 Woman, young, high cheekbones, smiling, broad forehead, blond hair.
 This man is young. He is clean-shaven, has bushy eyebrows and a broad forehead. He has an oval face and short black hair.	 Man, an oval face, no beard, broad forehead, short black hair.